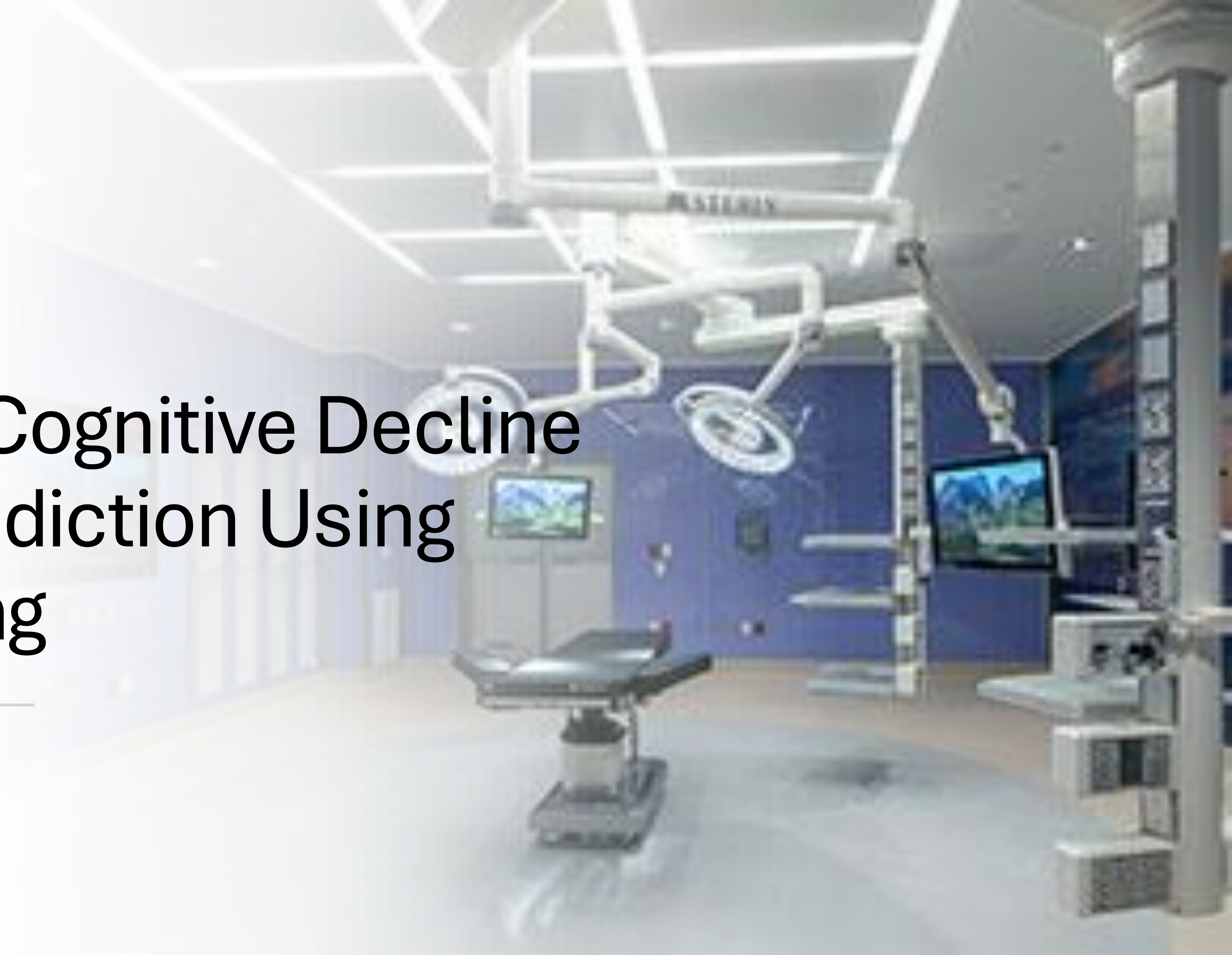
Post-Operative Cognitive Decline (POCD) Risk Prediction Using Machine Learning

-Vineeth Karjala

-Uttam Kumar Bellamkonda



Introduction

Some patients, especially older adults, may face **memory and thinking problems after surgery**. This condition is called POCD, and it can affect their ability to focus, remember things, or solve problems during recovery.

When POCD occurs, patients often **recover more slowly**, need more support, and may have to **stay longer in the hospital or go to rehabilitation** before returning home.

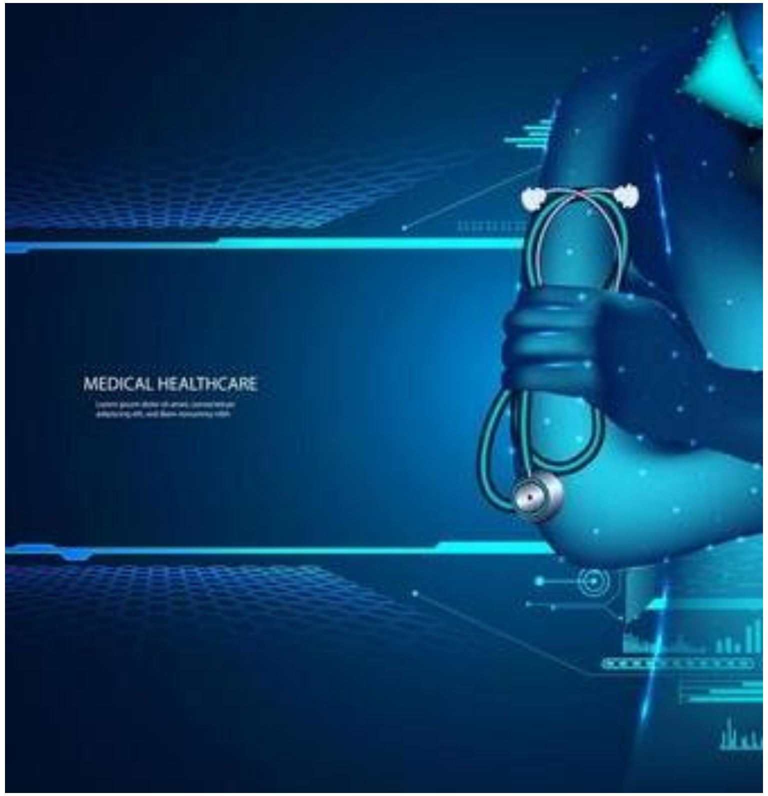
By using **machine learning models** trained on patient data, doctors can get an **early warning** about who is at high risk and provide better care.

Literature Review

- POCD affects up to 10–60% of older adults after major surgery, especially those requiring ICU care.
- Risk factors include advanced age, longer surgeries, ICU stay, comorbidities, hypotension, and anesthesia exposure.
- Recent studies highlight the need for early prediction to support clinical decision-making.
- Machine learning has shown promise in detecting high-risk postoperative patients using EHR data.

Problem Statement

- POCD is quite common, but it often goes undiagnosed or recognized late because there is no automatic clinical tool that can predict which patients are at higher risk after surgery.
- In most cases, doctors only identify POCD after symptoms start, such as confusion or memory issues — and by then, treatment and recovery become more challenging.
- By using hospital data and machine learning models, we aim to predict POCD in advance, helping doctors provide early monitoring and support to improve patient outcomes.



Project Goals

Build a machine learning model that uses **hospital records and surgery details** to identify patients who may have a **higher risk of developing POCD**.

Provide doctors with an **early warning**, so they can **monitor high-risk patients more closely** and support their cognitive recovery.

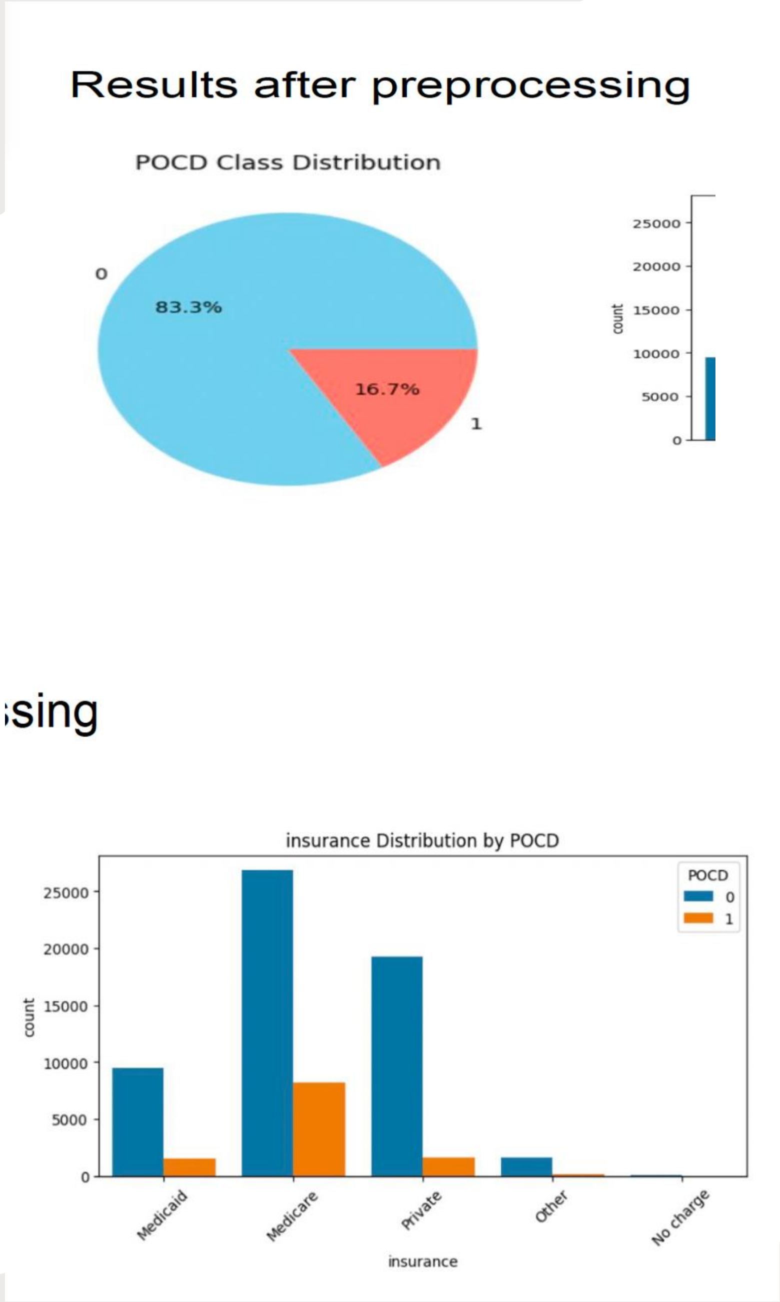
Create a system that can **improve safety and recovery**, especially for older adults, and can be **developed into a real clinical decision support tool** in the future.

Dataset

- We used the **MIMIC-IV v3.1** dataset from PhysioNet, which includes structured data such as patient demographics, surgery details, ICU stays, and diagnosis codes. Since this dataset does not include free-text clinical notes, we could not apply direct NLP methods.
- A separate resource, **MIMIC-IV-Note**, provides anonymized clinical notes for the same patients. In the future, combining both datasets can help build **multimodal machine learning models** and apply advanced NLP tools like **ClinicalBERT** to better detect cognitive changes.
- During preprocessing, we found duplicate entries and imbalance issues where most outcomes were labeled as no-POCD. After cleaning we obtained a balanced dataset suitable for building reliable models.
- Dataset link : https://physionet.org/content/mimiciv/3.1/
- We specifically focused on surgical ICU patients because this group has a significantly higher risk of developing POCD, making early prediction clinically impactful.

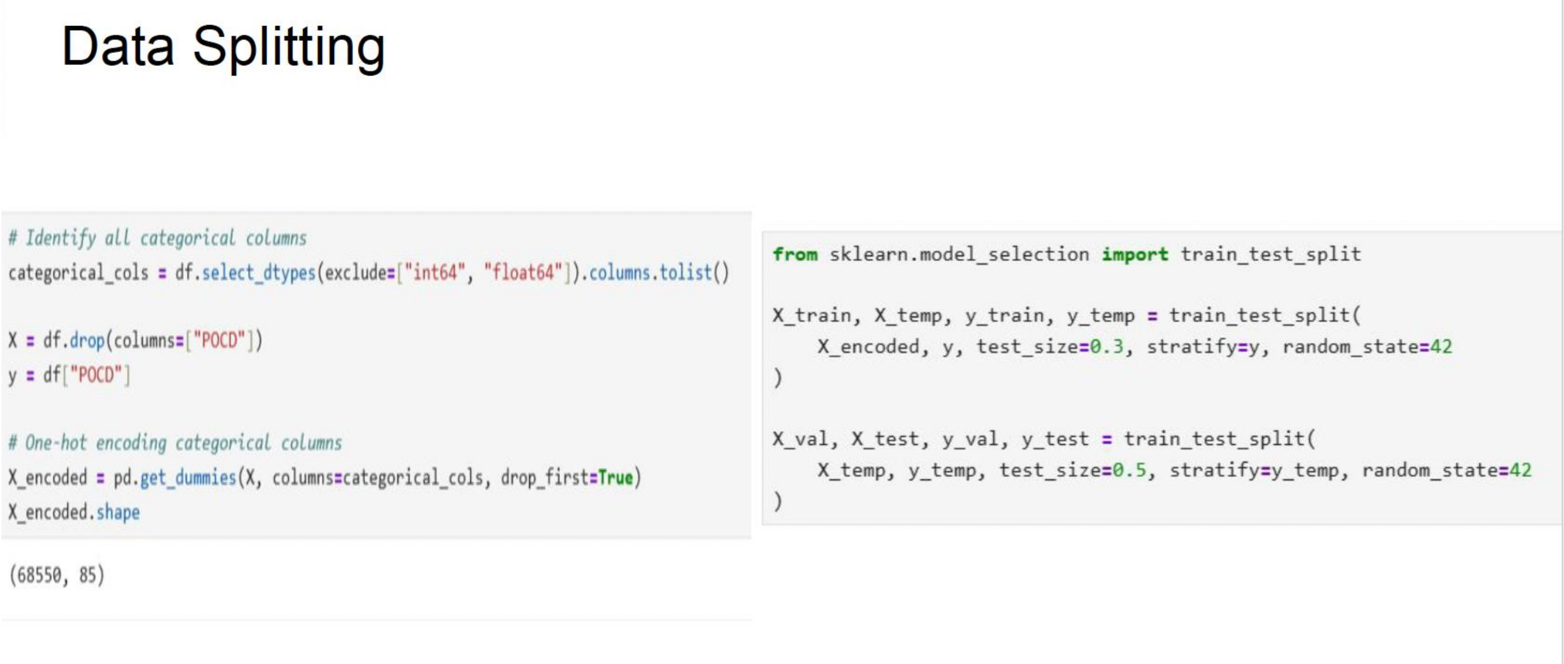
Data Processing Steps

- We combined important hospital records such as demographics, surgeries, ICU stays, and diagnosis codes, and **kept only patients who underwent a surgery** to focus on the POCD-risk group. We also created useful features like age, number of procedures, ICU transfers, and ICU stay duration, while **handling missing values** to maintain complete data.
- To ensure correct labeling, we **removed duplicate entries** and used ICD diagnosis codes to identify POCD and non-POCD patients accurately.
- Then split the data into training, validation, and testing sets using **stratified sampling** to keep class proportions consistent.



Results after preprocessing

Data Splitting



Data Splitting

```
# Identify all categorical columns
categorical_cols = df.select_dtypes(exclude=["int64", "float64"]).columns.tolist()

X = df.drop(columns=["POCD"])
y = df["POCD"]

# One-hot encoding categorical columns
X_encoded = pd.get_dummies(X, columns=categorical_cols, drop_first=True)
X_encoded.shape
```

```
(68550, 85)
```

```
from sklearn.model_selection import train_test_split

X_train, X_temp, y_train, y_temp = train_test_split(
    X_encoded, y, test_size=0.3, stratify=y, random_state=42
)

X_val, X_test, y_val, y_test = train_test_split(
    X_temp, y_temp, test_size=0.5, stratify=y_temp, random_state=42
)
```

- To build a reliable model, we divided the dataset into three parts: **60% for training**, where the model learns patterns from the data.

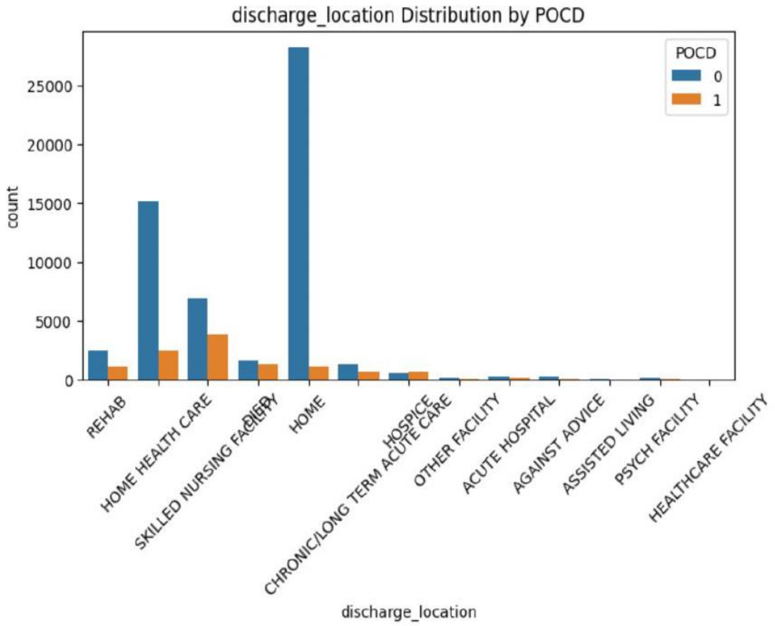
- **20% of the data** was used for **validation**, helping us adjust model parameters and improve performance during development.

- The remaining **20%** was kept aside for **final testing**, so we could evaluate how well the model works on completely unseen data. We used **stratified sampling** to keep the proportion of POCD and non-POCD patients **consistent** across all splits.

- We examined how POCD is related to the type of discharge patients receive after surgery, such as going home independently, with home care, or to a rehabilitation center.

- We observed that patients who needed more support after surgery (rehab, skilled nursing facilities, or home-care services) showed a higher number of POCD cases.

- This indicates that a patient’s functional recovery and support needs after surgery may be strongly linked to their risk of developing POCD.



Model Development

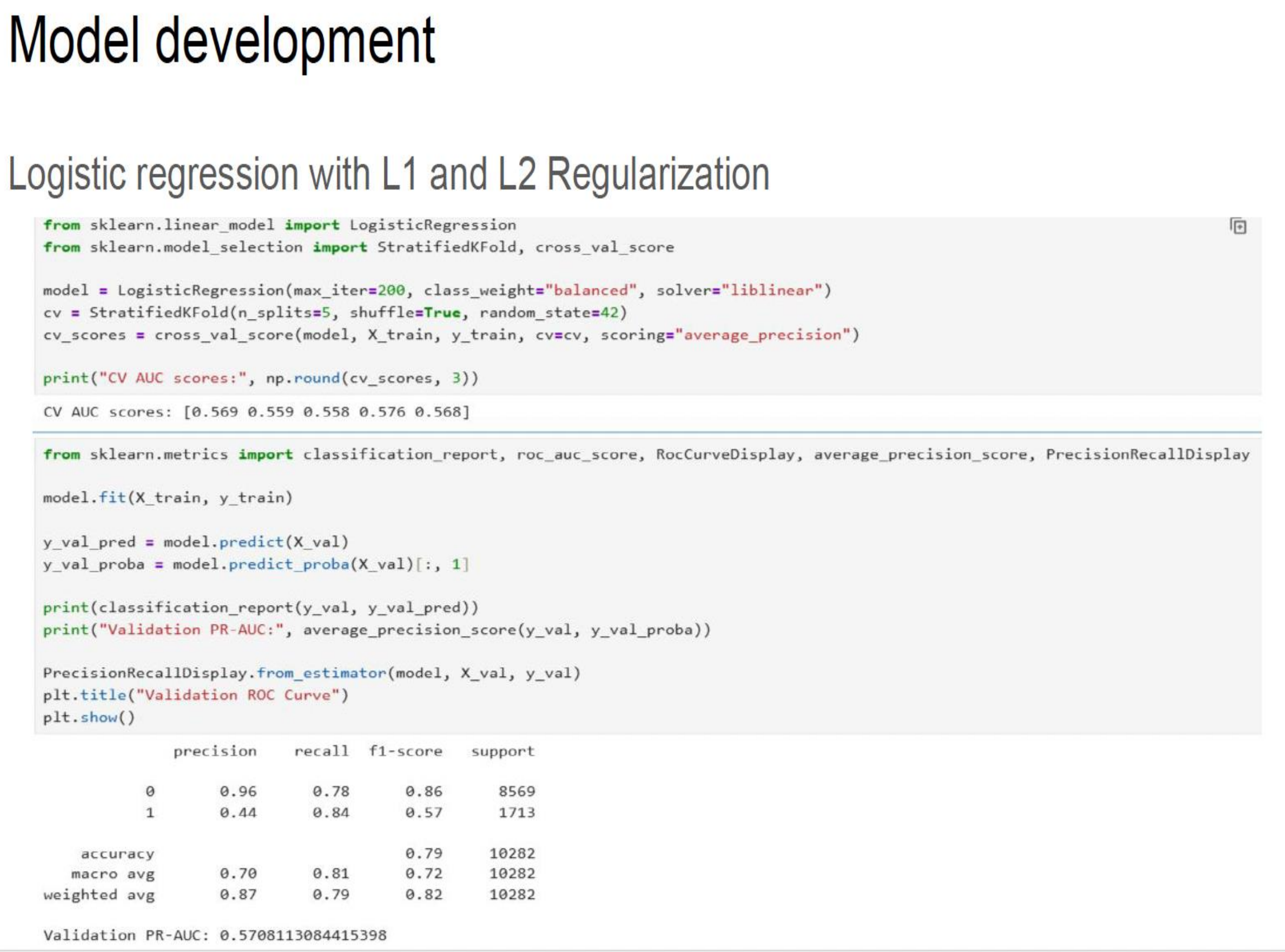
- We trained and compared **multiple machine learning models**, including Logistic Regression, Random Forest, and a Neural Network (MLP), to understand which performs best for POCD prediction.

- All models were trained using **stratified resampling** to preserve the same POCD proportion in each split.

- We tested the models on **unseen data** to measure how well they can predict POCD for new patients in real clinical situations.

Logistic Regression

- We first developed a Logistic Regression model because it is simple, fast, and easy to interpret in clinical settings.
- We used both **L1 and L2 regularization** to avoid overfitting and to understand which features are truly important.
- We trained the model on the balanced training set and evaluated it on unseen data.
- The validation PR-AUC was around **0.57**, showing a reasonable baseline for comparison.



Model development

Logistic regression with L1 and L2 Regularization

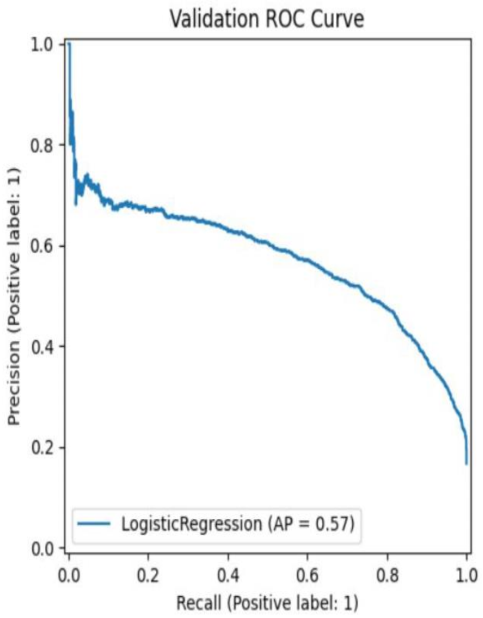
```
from sklearn.linear_model import LogisticRegression
from sklearn.model_selection import StratifiedKFold, cross_val_score

model = LogisticRegression(max_iter=200, class_weight="balanced", solver="liblinear")
cv = StratifiedKFold(n_splits=5, shuffle=True, random_state=42)
cv_scores = cross_val_score(model, X_train, y_train, cv=cv, scoring="average_precision")

print("CV AUC scores:", np.round(cv_scores, 3))
```

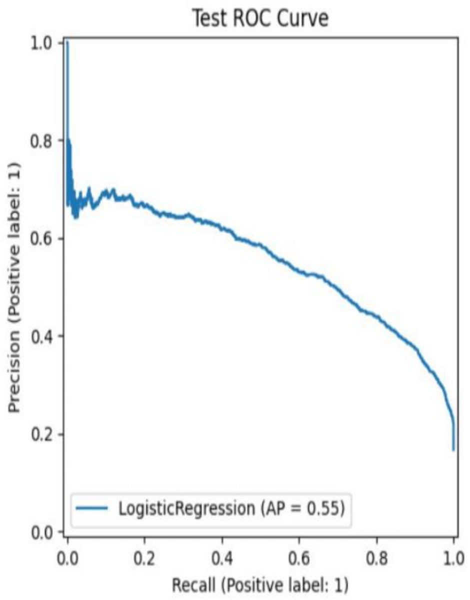
```
CV AUC scores: [0.569 0.559 0.558 0.576 0.568]
```

```
from sklearn.metrics import classification_report, roc_auc_score, RocCurveDisplay, average_precision_score, PrecisionRecallDisplay

model.fit(X_train, y_train)

y_val_pred = model.predict(X_val)
y_val_proba = model.predict_proba(X_val)[:, 1]

print(classification_report(y_val, y_val_pred))
print("Validation PR-AUC:", average_precision_score(y_val, y_val_proba))

PrecisionRecallDisplay.from_estimator(model, X_val, y_val)
plt.title("Validation ROC Curve")
plt.show()
```

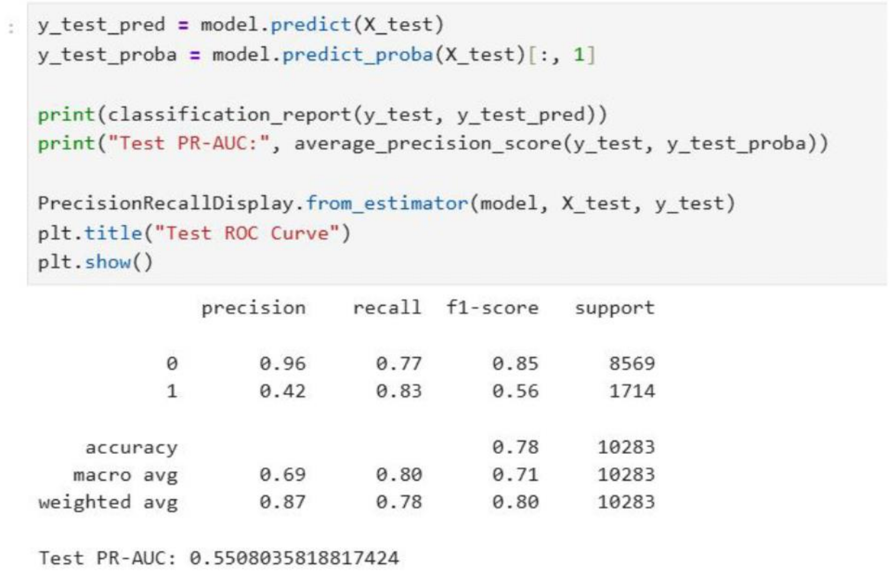
```
              precision    recall  f1-score   support

           0       0.96      0.78      0.86      8569
           1       0.44      0.84      0.57      1713

    accuracy                           0.79     10282
   macro avg       0.70      0.81      0.72     10282
weighted avg       0.87      0.79      0.82     10282

Validation PR-AUC: 0.5708113084415398
```

- We evaluated the model using Precision-Recall metrics.
  The model correctly identified many non-POCD cases,but recall for POCD cases was still limited due to the complexity of this clinical outcome.
- We applied GridSearchCV to find the best penalty, regularization strength, and solver.
  After tuning, the model improved slightly and became more stable.



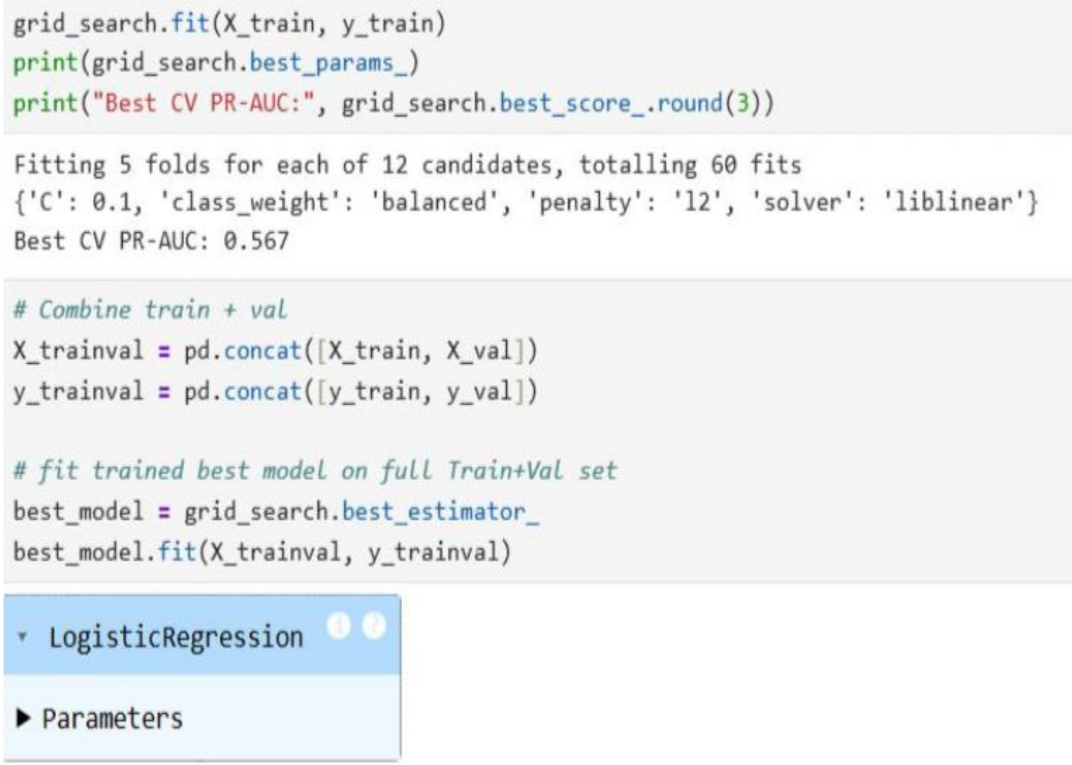


Test ROC



```
y_test_pred = model.predict(X_test)
y_test_proba = model.predict_proba(X_test)[:, 1]

print(classification_report(y_test, y_test_pred))
print("Test PR-AUC:", average_precision_score(y_test, y_test_proba))

PrecisionRecallDisplay.from_estimator(model, X_test, y_test)
plt.title("Test ROC Curve")
plt.show()
```

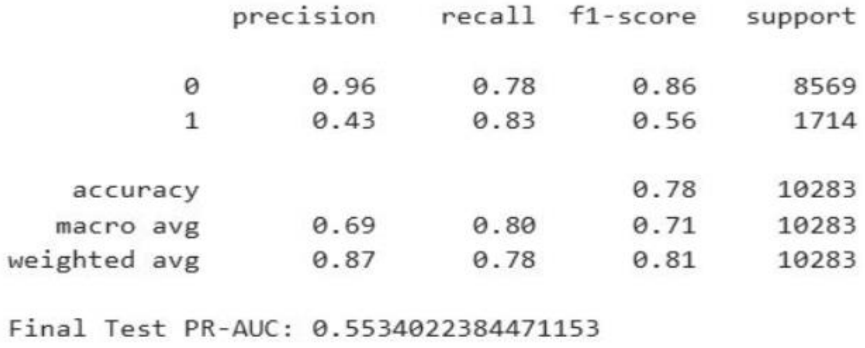
```
              precision    recall  f1-score   support

           0       0.96      0.77      0.85      8569
           1       0.42      0.83      0.56      1714

    accuracy                           0.78     10283
   macro avg       0.69      0.80      0.71     10283
weighted avg       0.87      0.78      0.80     10283

Test PR-AUC: 0.5508035818817424
```



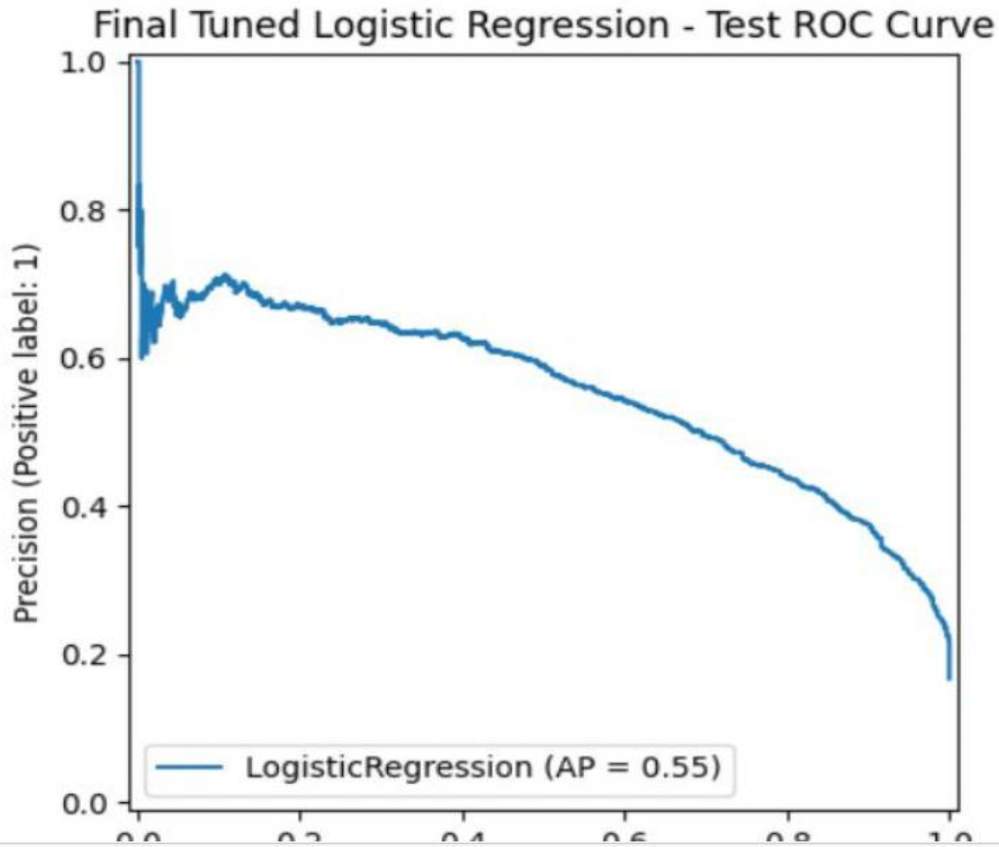
```
grid_search.fit(X_train, y_train)
print(grid_search.best_params_)
print("Best CV PR-AUC:", grid_search.best_score_.round(3))
```

```
Fitting 5 folds for each of 12 candidates, totalling 60 fits
{'C': 0.1, 'class_weight': 'balanced', 'penalty': 'l2', 'solver': 'liblinear'}
Best CV PR-AUC: 0.567
```

```
# Combine train + val
X_trainval = pd.concat([X_train, X_val])
y_trainval = pd.concat([y_train, y_val])

# fit trained best model on full Train+Val set
best_model = grid_search.best_estimator_
best_model.fit(X_trainval, y_trainval)
```

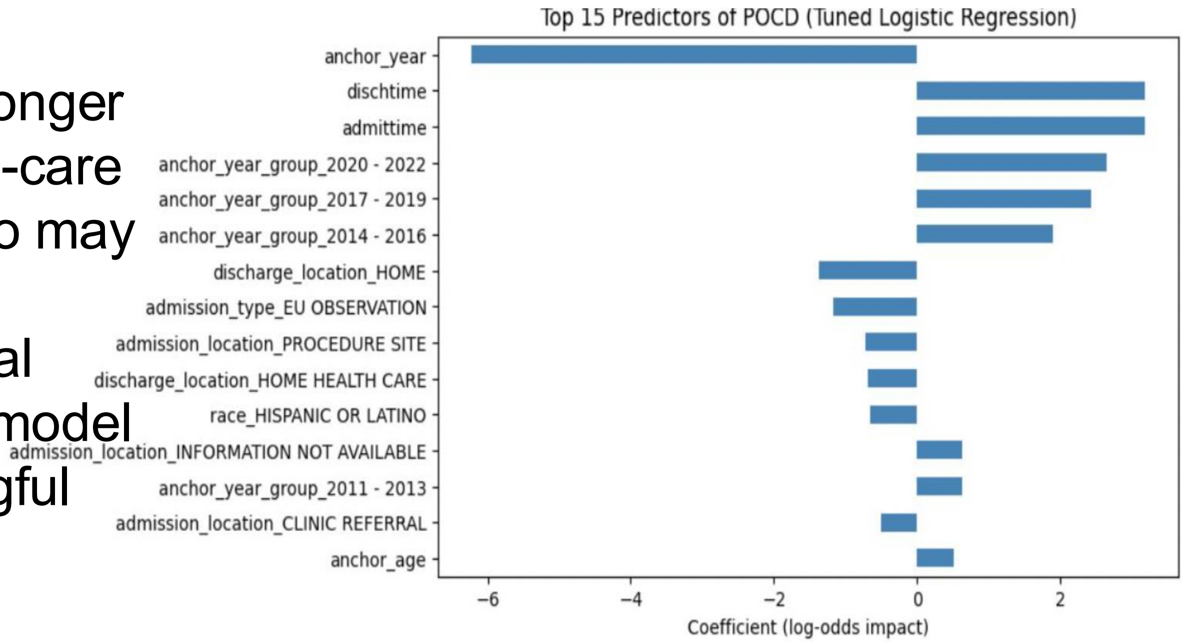


```
              precision    recall  f1-score   support

           0       0.96      0.78      0.86      8569
           1       0.43      0.83      0.56      1714

    accuracy                           0.78     10283
   macro avg       0.69      0.80      0.71     10283
weighted avg       0.87      0.78      0.81     10283

Final Test PR-AUC: 0.5534022384471153
```

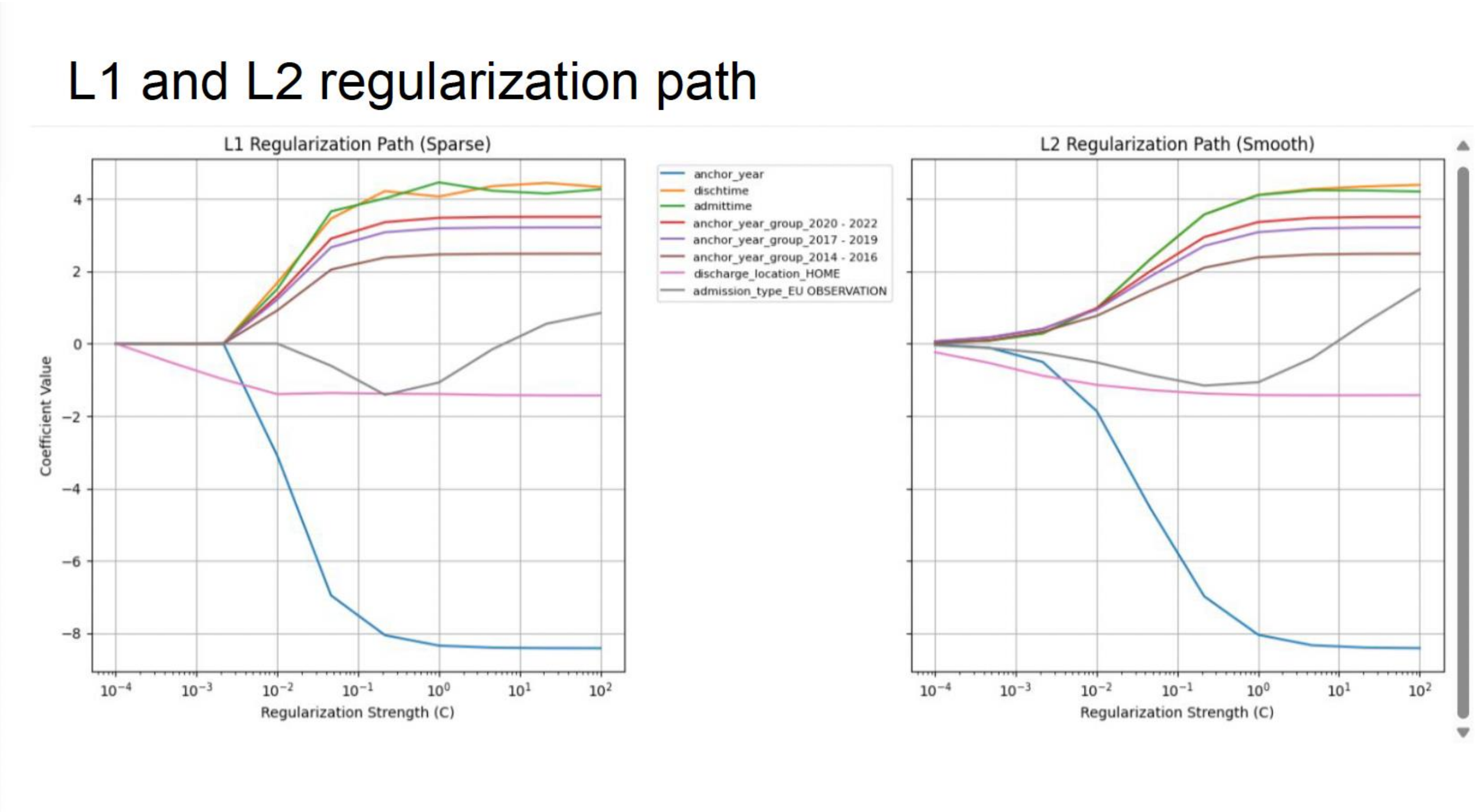


Important Predictors of POCD

- Logistic Regression helps us understand which clinical features contribute the most to POCD risk, making the results easier for doctors to interpret.
- Some of the strongest predictors included longer ICU stay, older age, and discharge to home-care or rehabilitation, which indicate patients who may need more support after surgery.
- These findings agree with real-world medical knowledge, supporting the reliability of our model and showing that it learns clinically meaningful patterns.



- We compared L1 (sparse) vs L2 (smooth) regularization paths to observe feature selection behavior.



L1 and L2 regularization path

Random Forest Model

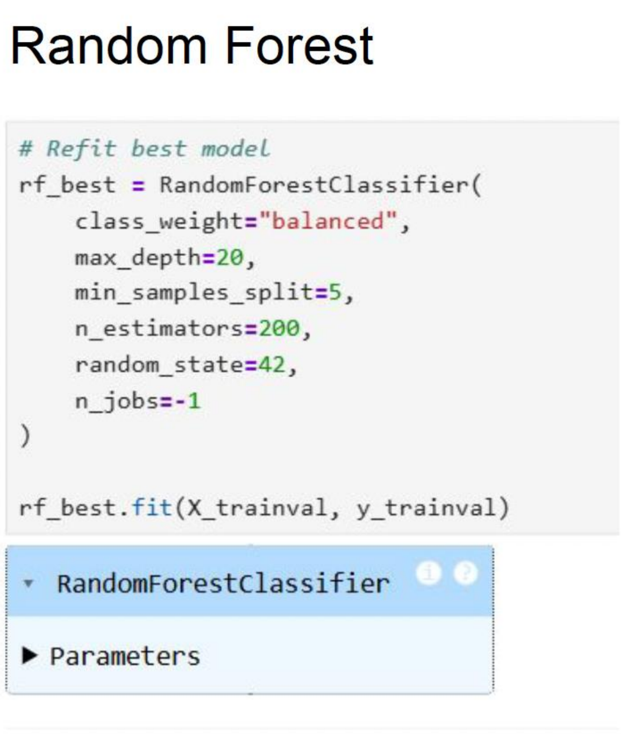
- Then, We used Random Forest because it can capture complex non-linear patterns in ICU data that Logistic Regression may miss.
- We trained the model using **stratified resampling**, so the POCD class proportion remains consistent in each train-test split.

We tuned key hyperparameters:

- Number of trees
- Tree depth

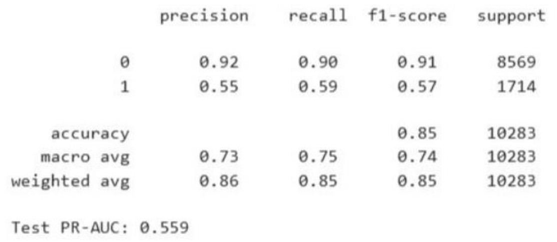
Minimum samples per split

- This model showed better recall for POCD cases.

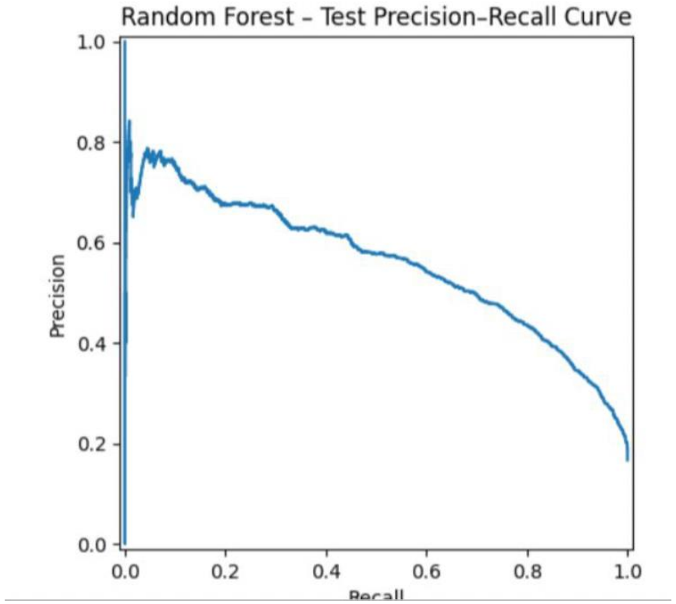


Random Forest

```
# Refit best model
rf_best = RandomForestClassifier(
    class_weight="balanced",
    max_depth=20,
    min_samples_split=5,
    n_estimators=200,
    random_state=42,
    n_jobs=-1
)

rf_best.fit(X_trainval, y_trainval)
```

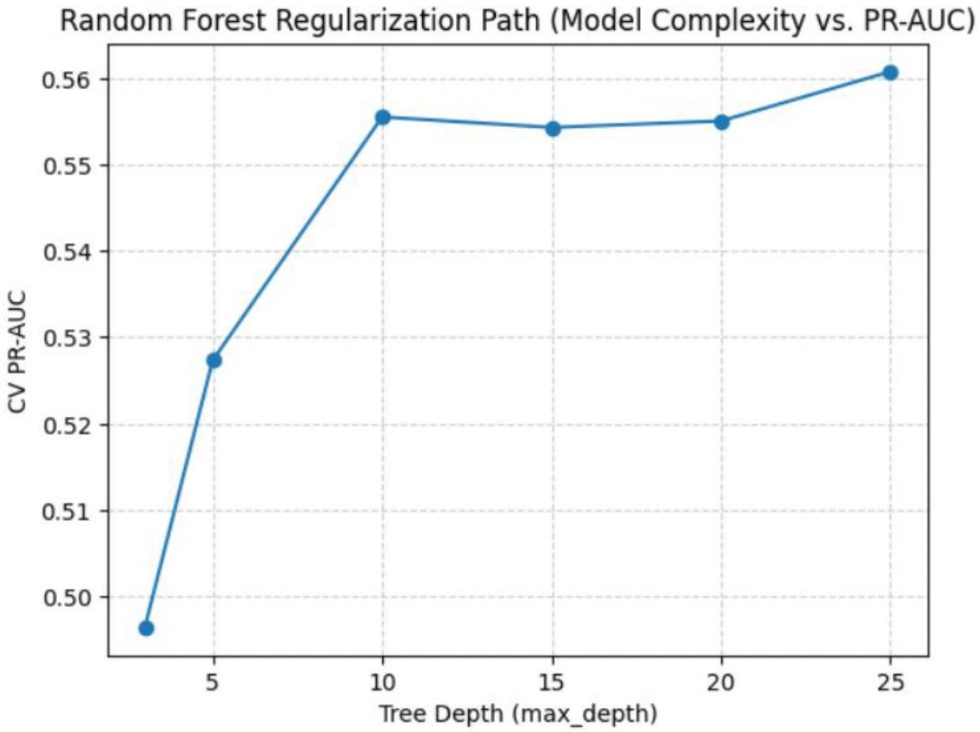


```
              precision    recall  f1-score   support

           0       0.92      0.90      0.91      8569
           1       0.55      0.59      0.57      1714

    accuracy                           0.85     10283
   macro avg       0.73      0.75      0.74     10283
weighted avg       0.86      0.85      0.85     10283

Test PR-AUC: 0.559
```



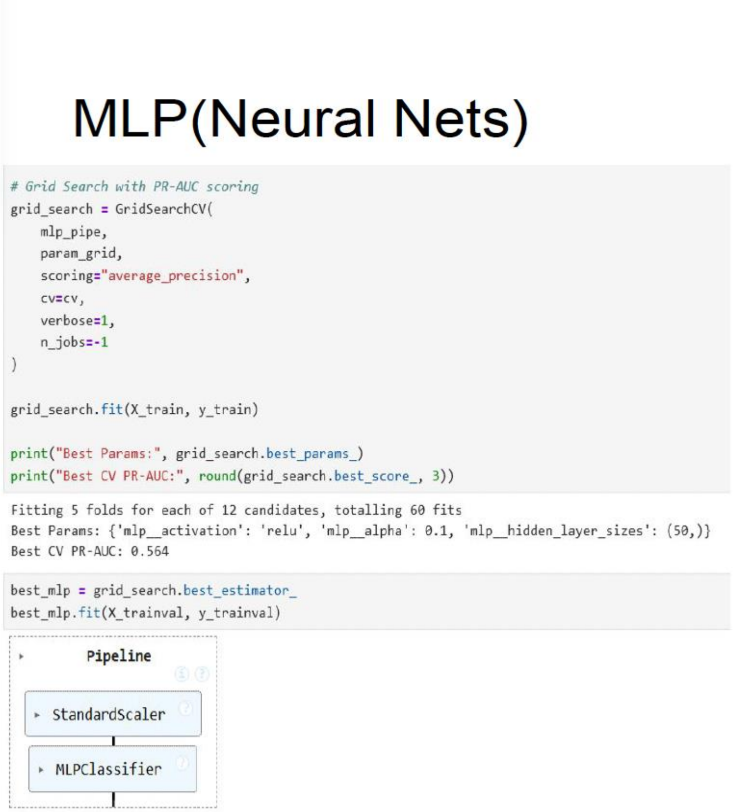
Random Forest – Test Precision–Recall Curve

- We tested different tree depths to find the best trade-off between accuracy and overfitting.



Neural Network (MLP) Model

- We used an MLP classifier to learn hidden relationships between features.
  We applied GridSearchCV to find the best number of neurons and activation function.
- This model achieved a PR-AUC similar to the Random Forest and showed that learning deeper patterns helps improve detection of minority cases.



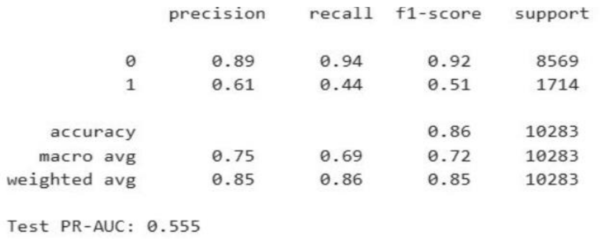
MLP(Neural Nets)

```
# Grid Search with PR-AUC scoring
grid_search = GridSearchCV(
    mlp_pipe,
    param_grid,
    scoring="average_precision",
    cv=cv,
    verbose=1,
    n_jobs=-1
)

grid_search.fit(X_train, y_train)

print("Best Params:", grid_search.best_params_)
print("Best CV PR-AUC:", round(grid_search.best_score_, 3))
```

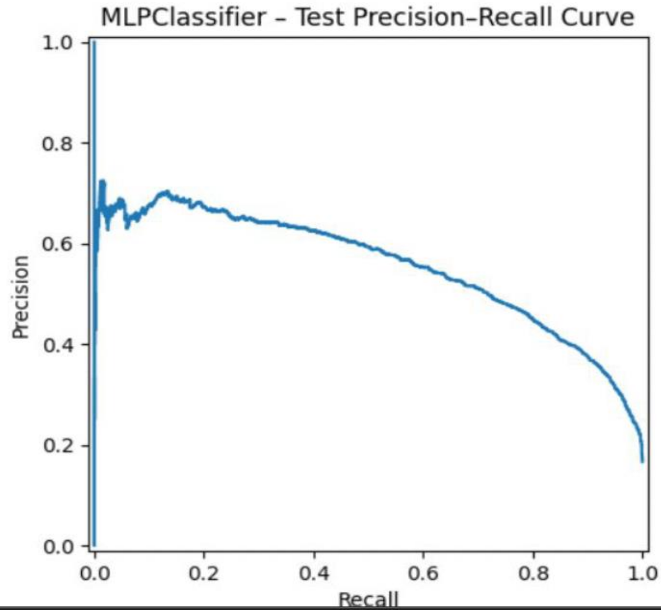
```
Fitting 5 folds for each of 12 candidates, totalling 60 fits
Best Params: {'mlp__activation': 'relu', 'mlp__alpha': 0.1, 'mlp__hidden_layer_sizes': (50,)}
Best CV PR-AUC: 0.564
```

```
best_mlp = grid_search.best_estimator_
best_mlp.fit(X_trainval, y_trainval)
```

Pipeline
StandardScaler
MLPClassifier



```
              precision    recall  f1-score   support

           0       0.89      0.94      0.92      8569
           1       0.61      0.44      0.51      1714

    accuracy                           0.86     10283
   macro avg       0.75      0.69      0.72     10283
weighted avg       0.85      0.86      0.85     10283

Test PR-AUC: 0.555
```



MLPClassifier – Test Precision–Recall Curve

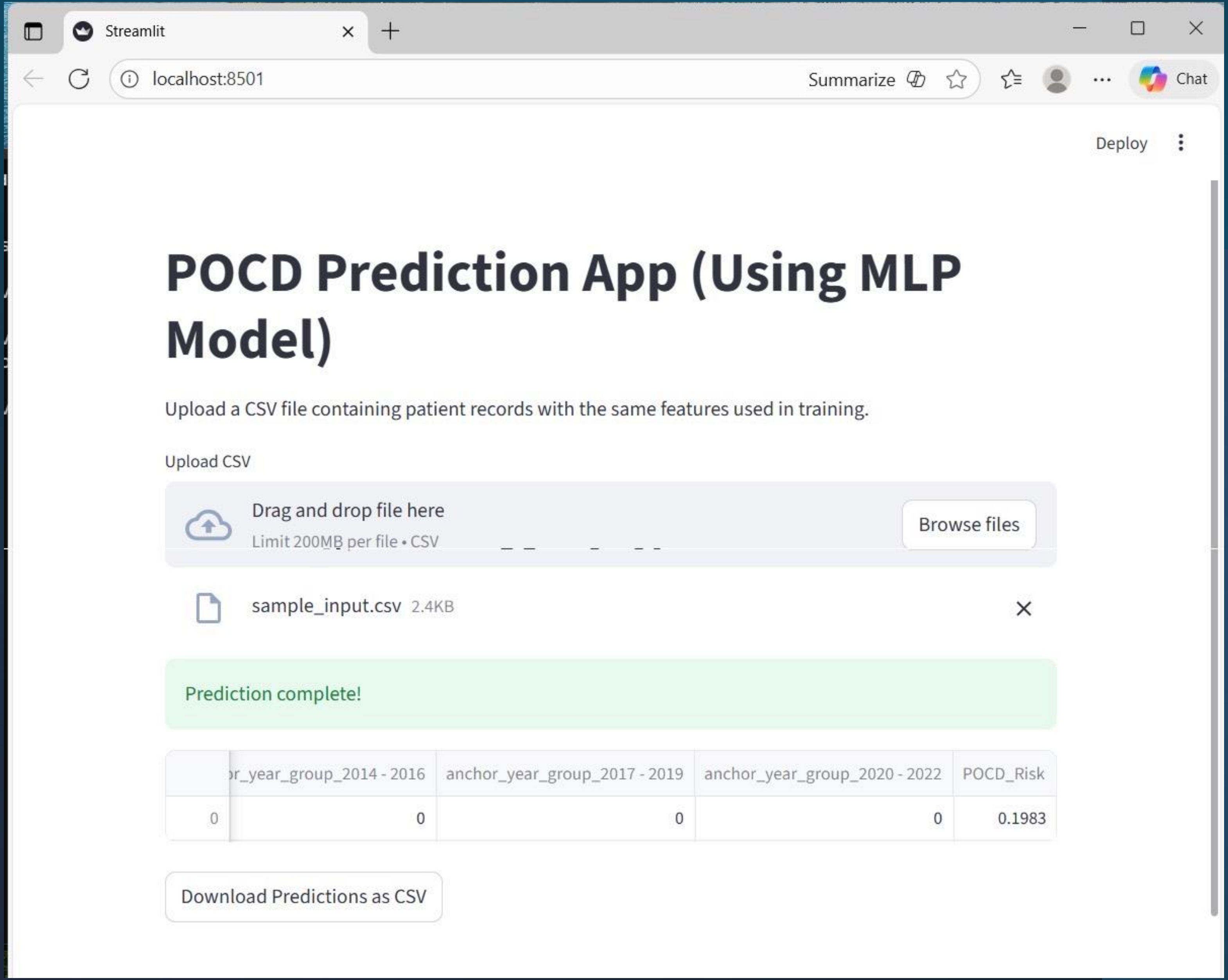
- Our models successfully detected POCD risk.
- Random Forest and MLP performed slightly stronger than Logistic Regression
- Key risk predictors such as:

Longer ICU stay

Need for discharge support

Older patient age
align well with known clinical risk patterns

- Results show that machine learning can support earlier POCD screening and better post-surgery care planning



POCD Prediction App (Using MLP Model)

Upload a CSV file containing patient records with the same features used in training.

Upload CSV

Drag and drop file here

Limit 200MB per file • CSV

sample_input.csv 2.4KB

Prediction complete!

	r_year_group_2014 - 2016	anchor_year_group_2017 - 2019	anchor_year_group_2020 - 2022	POCD_Risk
0	0	0	0	0.1983

Clinical Impact & Conclusion

- Helps doctors identify memory changes early instead of waiting for symptoms
- Early support may reduce long-term cognitive disability after surgery
- Better recovery and higher quality of life for elderly surgical patients
- Can lower healthcare costs by preventing complications
- Overall, machine learning can become a helpful screening tool to improve patient safety and post-operative care

Challenges & Limitations

VERY FEW POCD CASES IN DATA → MODELS STRUGGLE TO LEARN RARE EVENTS WELL

DATASET ONLY HAS STRUCTURED HOSPITAL DATA → NO SURGEON NOTES OR COGNITIVE TEST INFORMATION

POCD DIAGNOSIS IN RECORDS MAY NOT ALWAYS BE COMPLETE OR CONSISTENT

Future Scope

Including clinical notes from MIMIC-IV Notes can provide richer understanding of symptoms and cognitive condition

More advanced machine learning models (e.g., XGBoost, LightGBM) may improve prediction performance

Adding anesthesia-related factors, medications, and comorbid conditions can help capture additional clinical risks

References

- Johnson, A. E. W., Bulgarelli, L., Shen, L., et al. (2023). MIMIC-IV, a freely accessible electronic health record dataset. *Scientific Data, 10*(1), 1. https://doi.org/10.1038/s41597-022-01899-x
- Zhao, Q., Wan, H., Pan, H., & Xu, Y. (2024). Postoperative cognitive dysfunction—current research progress. *Frontiers in Behavioral Neuroscience, 18*, 1328790. https://doi.org/10.3389/fnbeh.2024.1328790
- Arefayne, N. R., Berhe, Y. W., & van Zundert, A. A. (2023). Incidence and factors related to prolonged postoperative cognitive decline in elderly patients following surgery and anaesthesia: A systematic review. *Journal of Multidisciplinary Healthcare, 16*, 3405–3413. https://doi.org/10.2147/JMDH.S431168
- Kotekar, N., Kuruvilla, C. S., & Murthy, V. (2014). Post-operative cognitive dysfunction in the elderly: A prospective clinical study. *Indian Journal of Anaesthesia, 58*(3), 263–268. https://doi.org/10.4103/0019-5049.135034
- Evered, L. A., & Silbert, B. S. (2018). Postoperative cognitive dysfunction and noncardiac surgery. *Anesthesia & Analgesia, 127*(2), 496–505. https://doi.org/10.1213/ANE.0000000000003514
- Borchers, F., Spies, C. D., Feinkohl, I., et al. (2021). Methodology of measuring postoperative cognitive dysfunction: A systematic review. *British Journal of Anaesthesia, 126*(6), 1119–1127. https://doi.org/10.1016/j.bja.2021.01.035
- Anand, N., Gupta, R., Mishra, S. P., & Mishra, M. (2024). Postoperative cognitive dysfunction: A review. *Asian Journal of Anesthesiology, 62*(1), 1–11. https://doi.org/10.6859/aja.202403_62(1).0001
- Wu, Y., Yu, C., & Gao, F. (2023). Risk factors for postoperative cognitive dysfunction in elderly patients undergoing surgery for oral malignancies. *Perioperative Medicine, 12*(42), 1–9. https://doi.org/10.1186/s13741-023-00330-2
- Sun, J., Du, X., & Chen, Y. (2024). Current progress on postoperative cognitive dysfunction: An update. *Journal of Integrative Neuroscience, 23*(12), 224. https://doi.org/10.31083/j.jin2312224

Thank you